# Storage Resource Management for Data Grid Applications

Arie Shoshani, LBNL
Don Petravick, Fermilab

Additional Staff

LBNL: Junmin Gu, Alex Sim, Alex Romosan (PT), Viji Natarajan (PT)
Fermilab: Timur Perelmutov

http://sdm.lbl.gov/srm

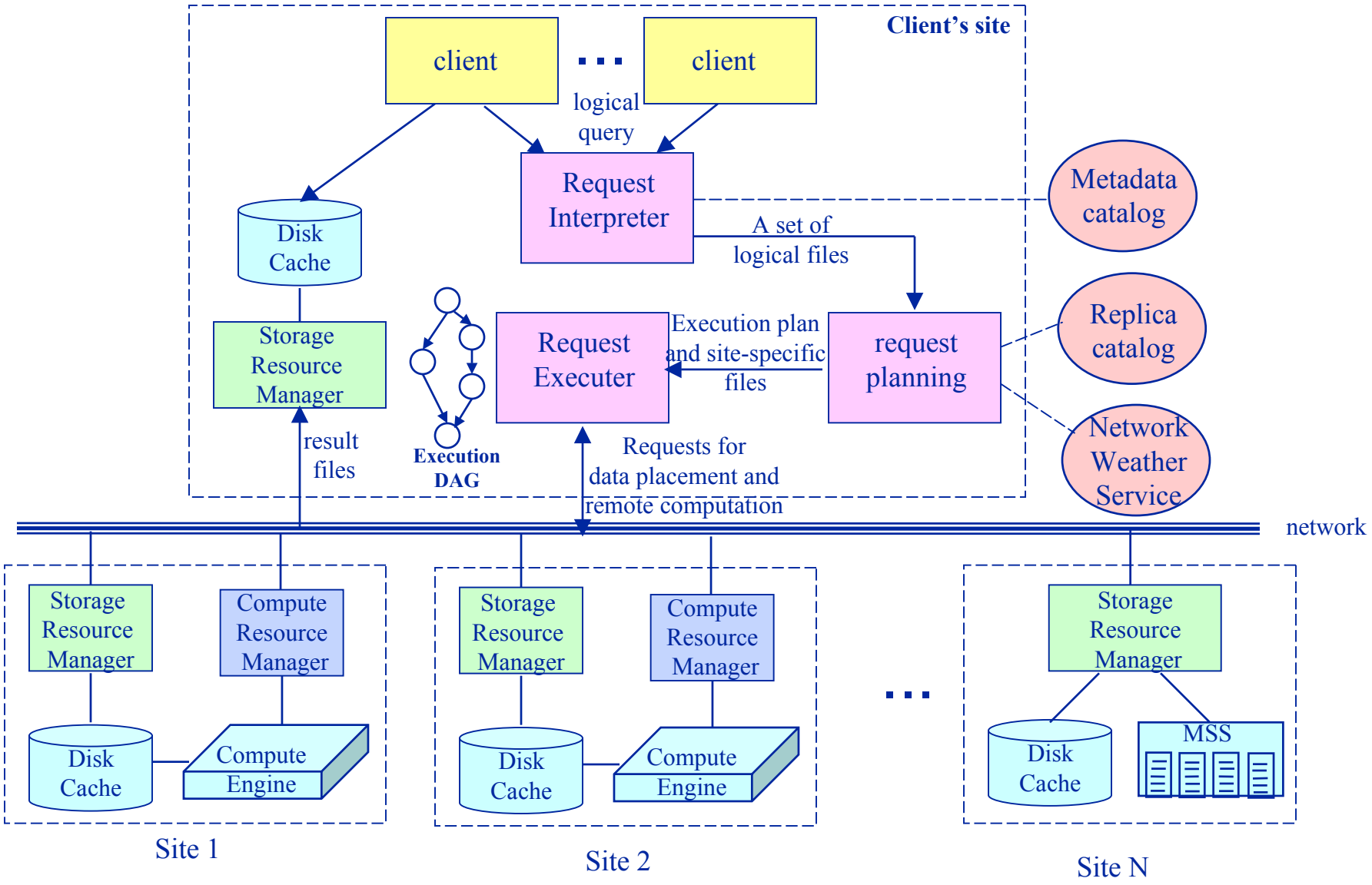
# Outline

- What are Storage Resource Managers – Objectives
- Example analysis scenario and the use of SRMs
- SRM challenges – functionality for various scenarios
- Progress - deployed SRMs in production
- Progress - standardization in US and EU
- Future plans

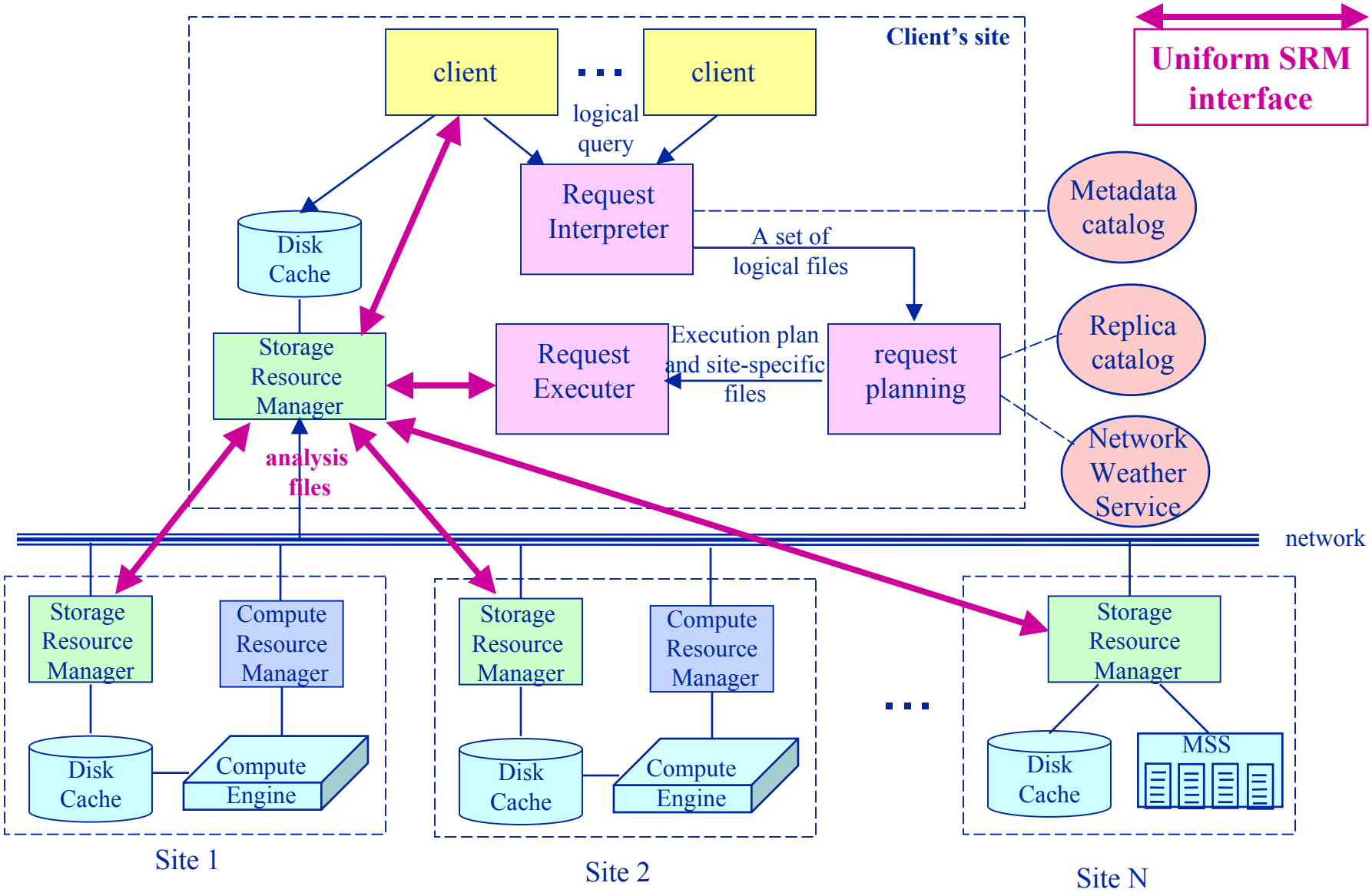
# Motivation

- Grid architecture needs to include reservation & scheduling of:
  - Compute resources
  - Storage resources
  - Network resources
- Storage Resource Managers (SRMs) role in the data grid architecture
  - Shared storage resource allocation & scheduling
  - Especially important for data intensive applications
  - Often files are archived on a mass storage system (MSS)
  - Wide area networks – minimize transfers
  - large scientific collaborations (100's of nodes, 1000's of clients) – opportunities for file sharing
  - File replication and caching may be used
  - Need to support non-blocking (asynchronous) requests

# General Analysis Scenario



# Bring-to-Local Analysis Scenario



# Challenges: SRM is a Service



- **SRM functionality**
  - **Manage space**
    - Negotiate and assign space to users
    - Manage "lifetime" of spaces
  - **Manage files on behalf of a user**
    - Pin files in storage till they are released
    - Manage "lifetime" of files
    - Manage action when pins expire (depends on file types)
  - **Manage file sharing**
    - Policies on what should reside on a storage resource at any one time
    - Policies on what to evict when space is needed
  - **Get files from remote locations when necessary**
    - Purpose: to simplify client's task
  - **Manage multi-file requests**
    - A brokering function: queue file requests, pre-stage when possible
  - **Provide grid access to/from mass storage systems**
    - HPSS (LBNL, ORNL, BNL), Enstore (Fermi), JasMINE (Jlab), Castor (CERN), MSS (NCAR), …
  - **Transfer protocol negotiation**
    - Gridftp, FTP, HTTP, bbftp, …

# SRM Methods



**File Movement**
srm(Prepare)Get:
srm(Prepare)Put:
srmCopy:

**Lifetime management**
srmReleaseFiles:
srmPutDone:
srmExtendFileLifeTime:

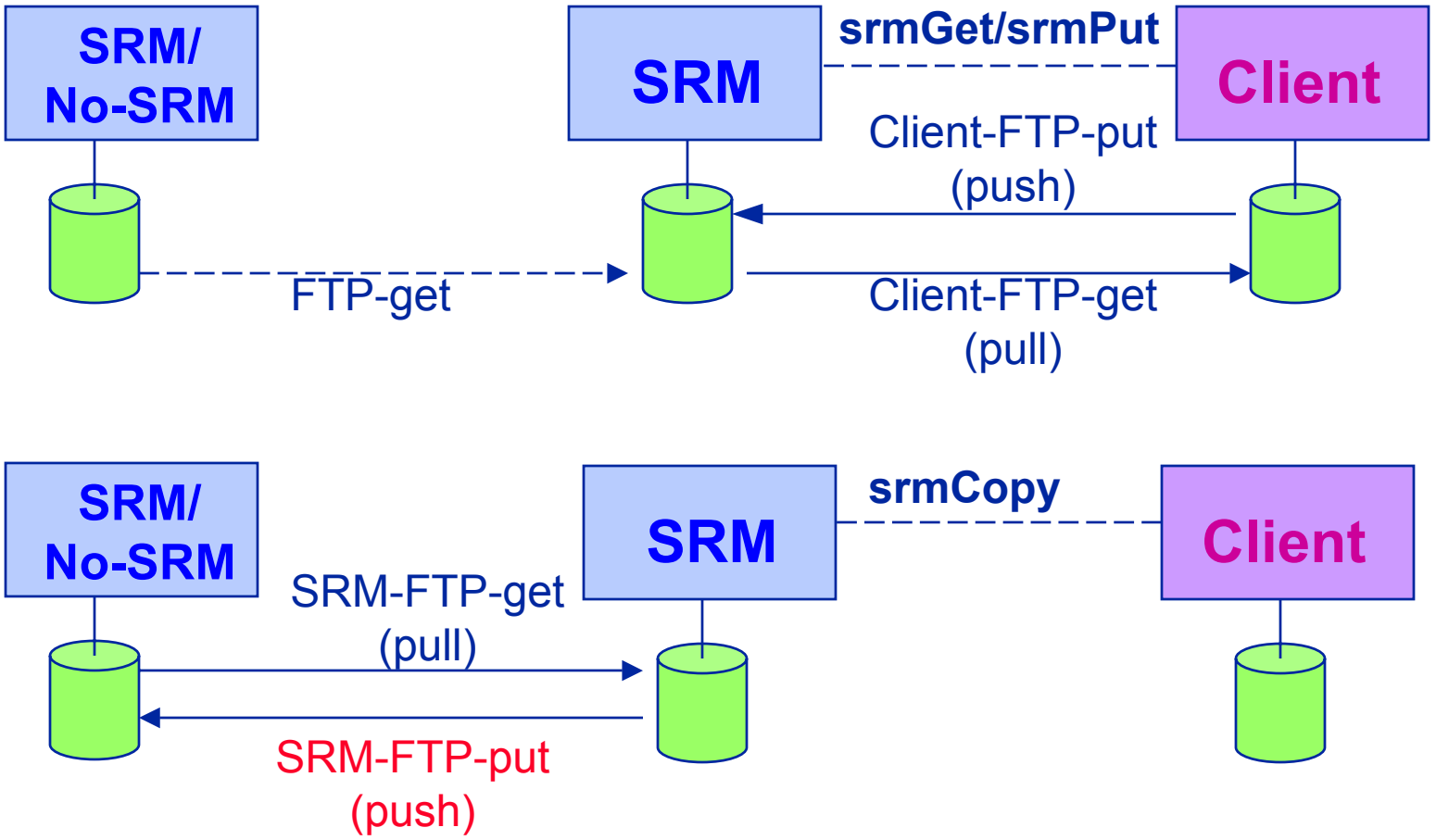
**Terminate/resume**
srmAbortRequest:
srmAbortFile
srmSuspendRequest:
srmResumeRequest:

**Space management**
srmReserveSpace
srmReleaseSpace
srmUpdateSpace
srmCompactSpace:
srmGetCurrentSpace:

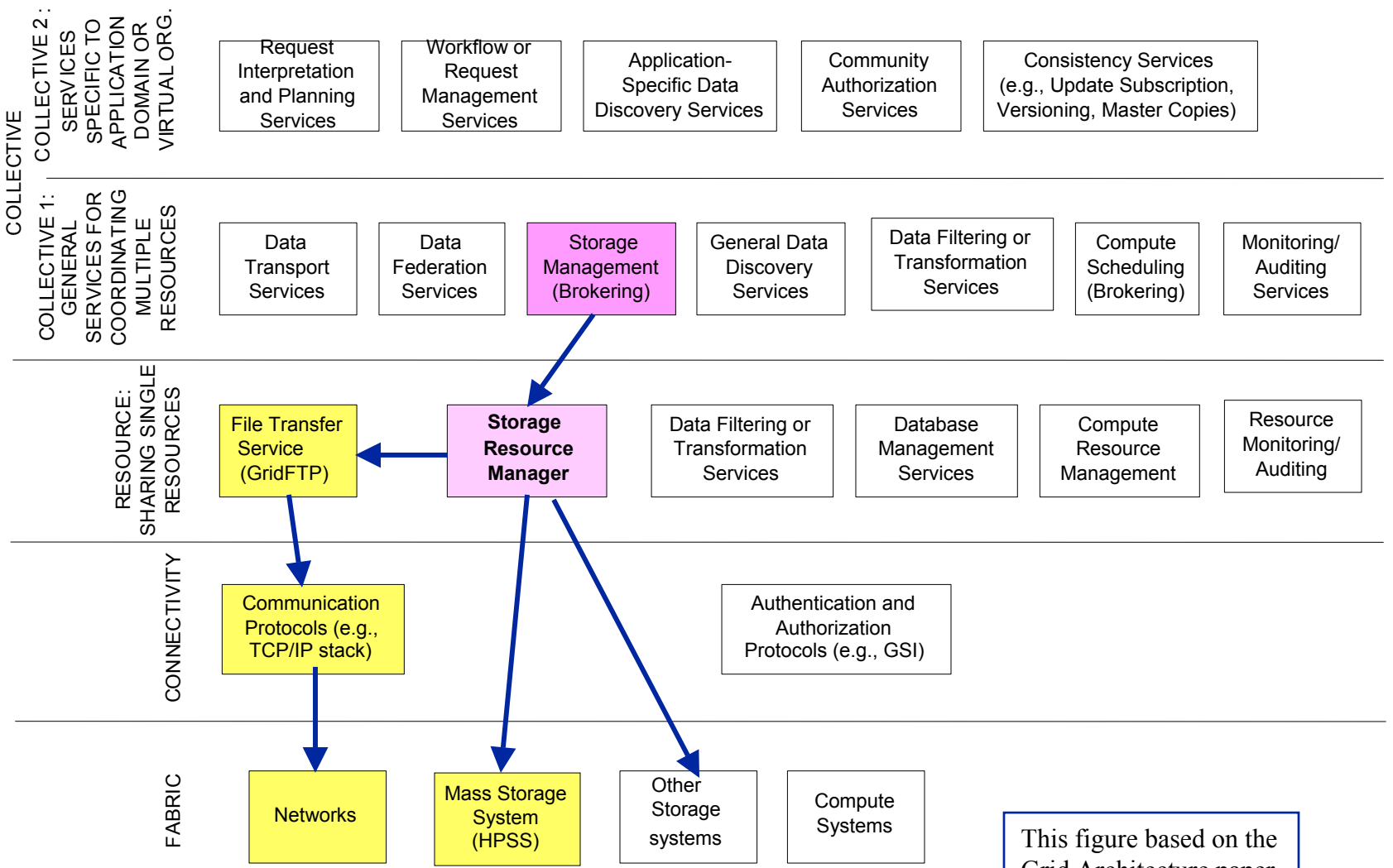
**FileType management**
srmChangeFileType:

**Status/metadata**
srmGetRequestStatus:
srmGetFileStatus:
srmGetRequestSummary:
srmGetRequestID:
srmGetFilesMetaData:
srmGetSpaceMetaData:

# File movement functionality: srmGet, srmPut, srmCopy



# SRMs provide a brokering service by supporting multi-file requests




This figure based on the Grid Architecture paper by Globus Team

# Types of SRMs



- **Types of storage resource managers**
  - Disk Resource Manager (DRM)
    - Manages one or more disk resources
  - Tape Resource Manager (TRM)
    - Manages access to a tertiary storage system (e.g. HPSS)
  - Hierarchical Resource Manager (HRM=TRM + DRM)
    - An SRM that stages files from tertiary storage into its disk cache

- **SRMs and File transfers**
  - SRMs DO NOT perform file transfer
  - SRMs DO invoke file transfer service if needed (GridFTP, FTP, HTTP, …)
  - SRMs DO monitor transfers and recover from failures
    - TRM: from/to MSS
    - DRM: from/to network

# Progress-to-Date

- **Interoperability**
  - SRM V1.0 adopted and implemented at three US and two EU institutions (LBNL, Fermilab, TJNAF, CERN-Castor, Rutherford Appleton Lab)
  - SRM V2.0 designed jointly by US and EU (space reservations, directory management)
  - Presented concepts at GGF
- **Prototypes**
  - Analysis scenario (SC 2001)
  - Interoperability (SC 2002)
  - ESG analysis (SC 2002)
- **Deployment**
  - PPDG-STAR Experiment – Robust File Replication BNL-NERSC (HPSS)
  - ESG project – adaptation to NCAR-MSS
  - ESG project – Robust File Replication NCAR-ORNL-LBNL
  - Web-based File Monitoring Tool
  - PPDG – Fermilab (CDF, US/CMS)
  - Lattice QCD grid – Fermilab, TJNAF, BNL (accessing NERSC, NCSA)

# SRM works with disk caches as well as legacy systems

LBNL

Fermilab

**SC 2001 Demo For HENP – STAR Experiment**

Logical Request

Denver

client

BIT-MAP Index

Request Manager

File Transfer Monitoring

DRM

GridFTP

Disk Cache

Legend:

Control path

Data Path

Chicago

server

GridFTP

Disk Cache

Berkeley

server

GridFTP

DRM

Disk Cache

Livermore

server

FTP

Disk Cache

Berkeley

server

GridFTP

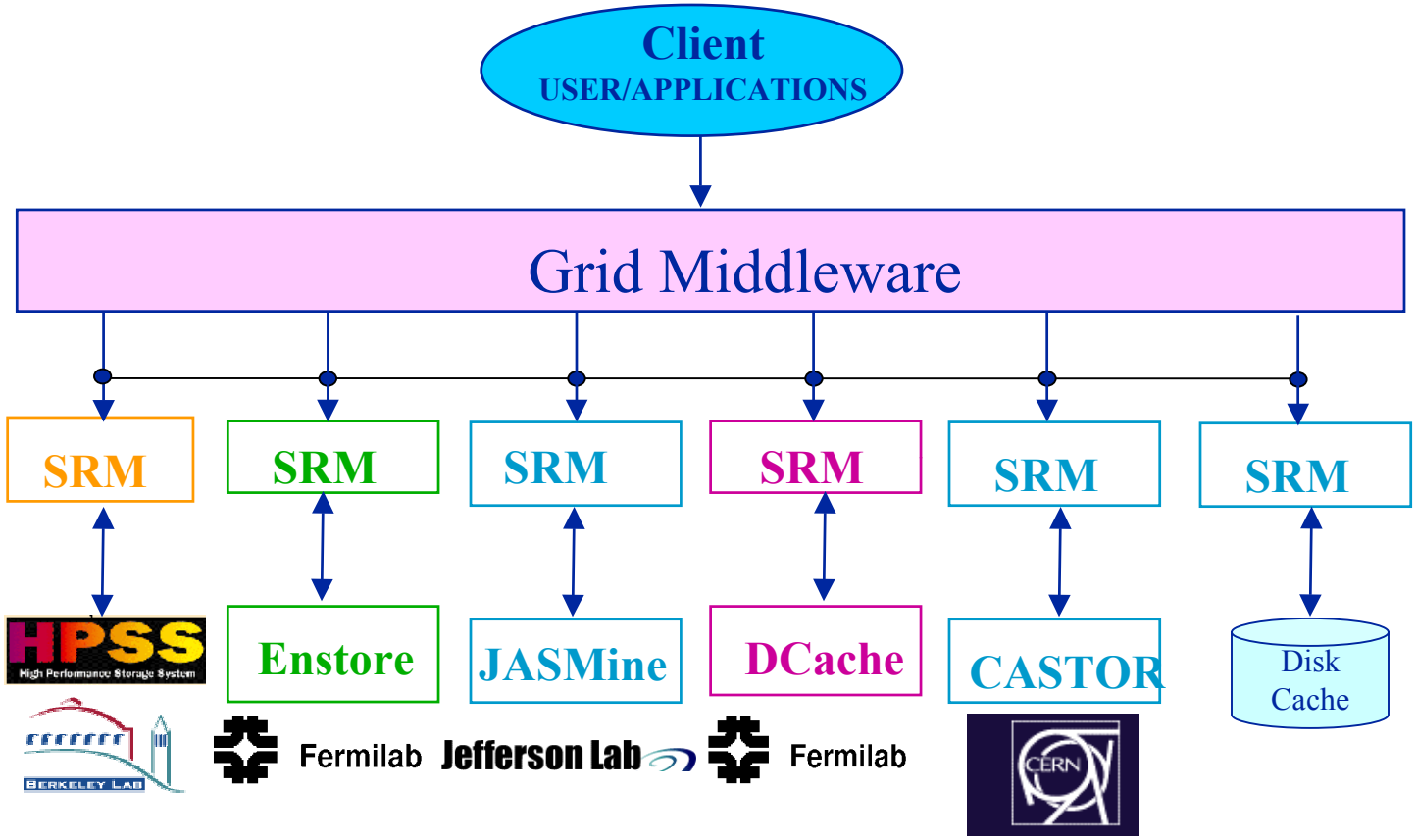
HRM

Disk Cache

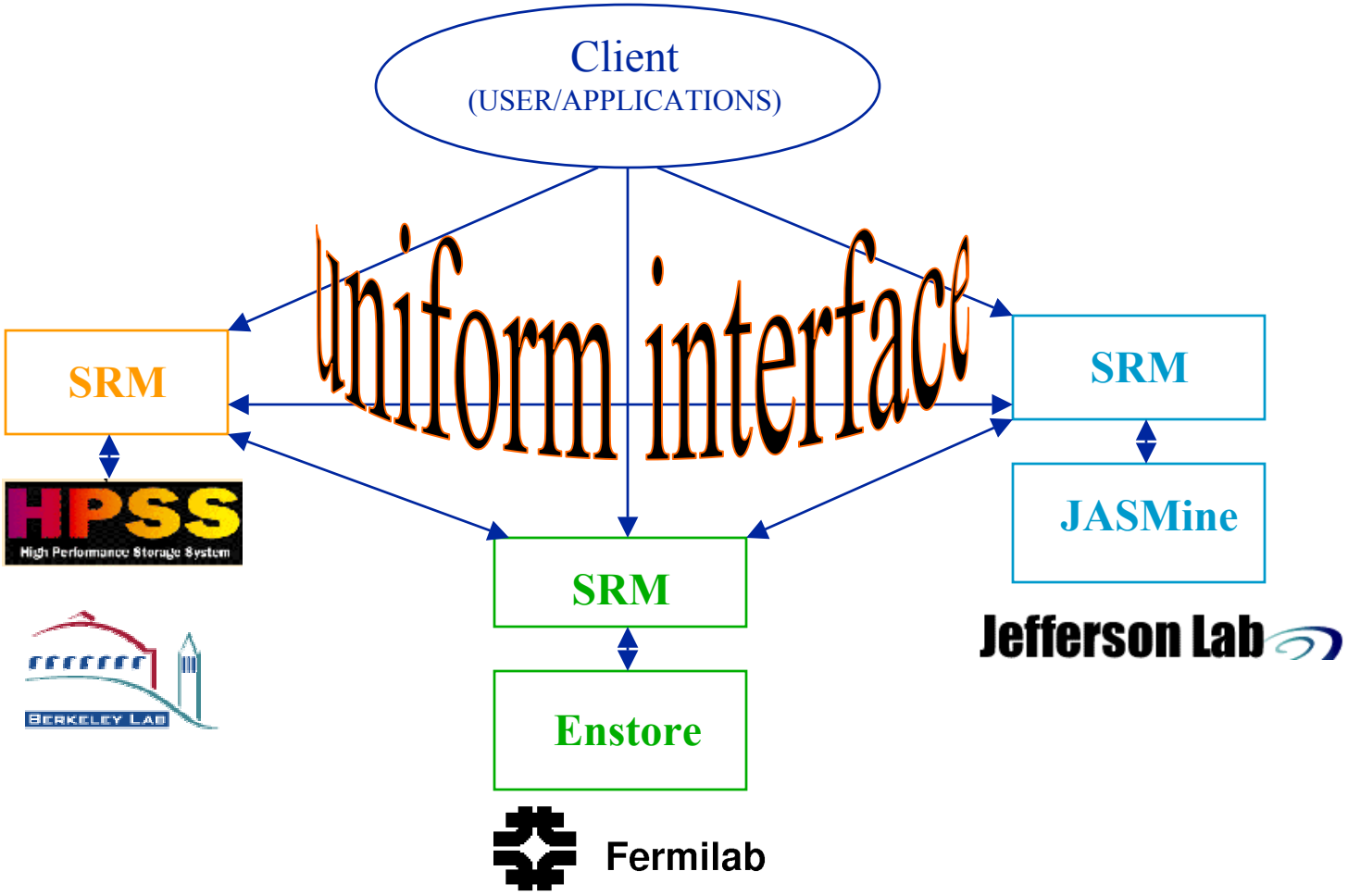
HPSS

High Performance Storage System

# Uniformity of Interface → Compatibility of SRMs



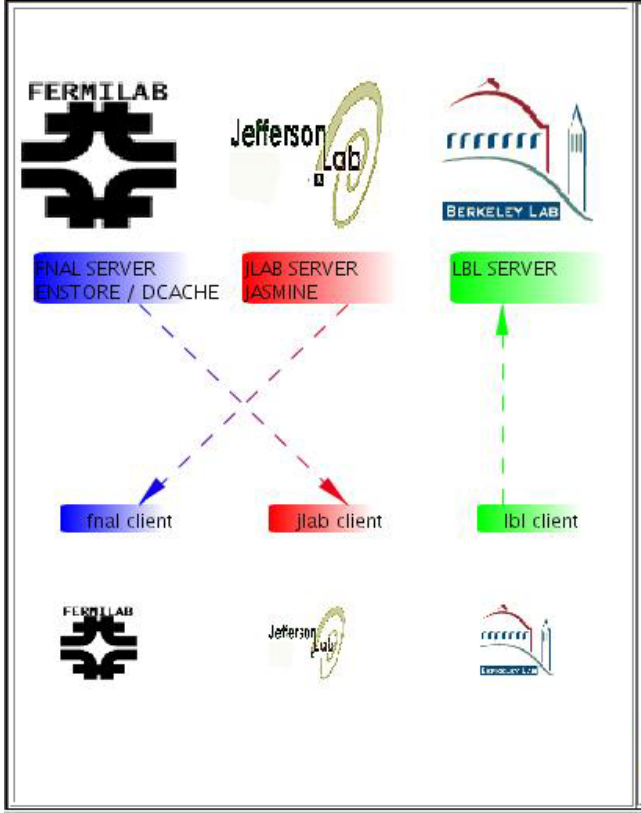
# High Level View of SRM setup in SC 2002



# Screen Dump of Demo at Fermi Booth

Fermilab

## Storage Resource Manager




### Storage Resource Manager

**Uniform Access to Mass Storage Systems**

HPSS, DCache/Enstore, Jasmine

**Provides**

- **Grid view of Site Storage Resources**
  (Site URL (SURL) )
- **Support for Local Policy**
  (user priviledges, bandwidth limits, robot sharing)
- **Advanced Storage Resource Allocation**
  (prestaging , space allocation, pinning)
- **Transfer Protocol Negotiation**
  (Site URL to Transfer URL (TURL) translation)

BERKELEY LAB FERMILAB Jefferson Lab

```
connecting to server
connecting to https://grid2.jlab.org:8443/glue/urn:hrm.wsdl
***                                                    ***
***          Welcome to the IAIK SSL (iSaSiLK) Library
***                                                    ***
*** This version of iSaSiLk is licensed for educational and resea
*** and evaluation only. Commercial use of this software is pro
*** For details please see http://jcewww.iaik.at/legal/license.h
*** This message does not appear in the registered commerci
***                                                    ***

connected to server, obtaining proxy
 install SslGsiSocketFactory as ssl and tcp factory
got proxy of type class $Proxy0
```

```
[15:10:35] Contacted server at httpg://grid2.jlab.org:8443/g
[15:10:35] Authorization not set
[15:10:45] [514][1] Ready Get srm:/mss/home/bhess/9840/
[15:10:45] (+) 1 network copies
[15:10:45] [514][1] Get Running http://grid2.jlab.org:8080/x
[15:10:51] /home/timur/cvs/sc2002srm/data/jlb_sc2002_file
[15:10:51] [514][1] Get Done http://grid2.jlab.org:8080/xfer
[15:10:51] (-) 0 network copies
[15:10:55] [514][1] Done Get srm:/mss/home/bhess/9840/c
----------------------------------------------------------
$/home/timur/cvs/sc2002srm/scripts/../clients/jlab-srm/bin
[15:10:57] using SRM at httpg://grid2.jlab.org:8443/glue/urn
***                                                    ***
***          Welcome to the IAIK ICE Library
```

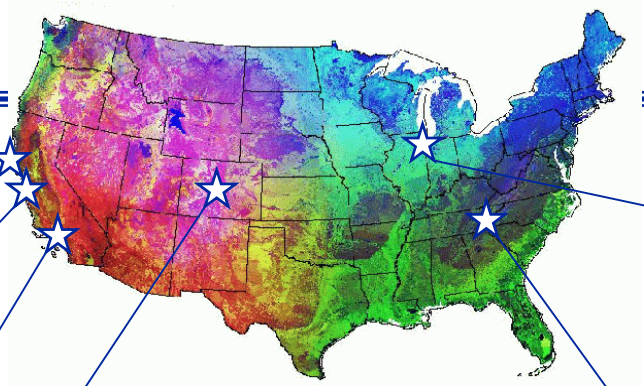
```
tcp buffer size: 1000000
num streams: 4
debug: 8
.. server replied with state = [Pending]
... from soap server, TURL = gsiftp://srm.lbl.gov/tmp/asim/2
----------------------------------------------------------

$/home/timur/cvs/sc2002srm/scripts/../clients/lbl-srm/hrm-
host: srm.lbl.gov
port: 8004
block size: 1000000
tcp buffer size: 1000000
num streams: 4
debug: 8
```

# Earth Science Grid Demo - SC 2002

LBNL

Fermilab



LBNL

HPSS
High Performance
Storage System

disk

HRM
Storage Resource
Management

gridFTP
server

LLNL

disk

DRM
Storage Resource
Management

gridFTP
server

USC-ISI

MCS
Metadata Cataloguing Services

SOAP

RLS
Replica Location Services

RMI

NCAR

openDAPg
server

gridFTP
Striped
server

Tomcat servlet engine

MCS client

MyProxy client

RLS client

DRM
Storage Resource
Management

MyProxy
server

CAS client

GRAM
gatekeeper

gridFTP

gridFTP
server

HRM
Storage Resource
Management

disk

NCAR-MSS
Mass Storage System

ANL

CAS
Community Authorization Services

ORNL

gridFTP

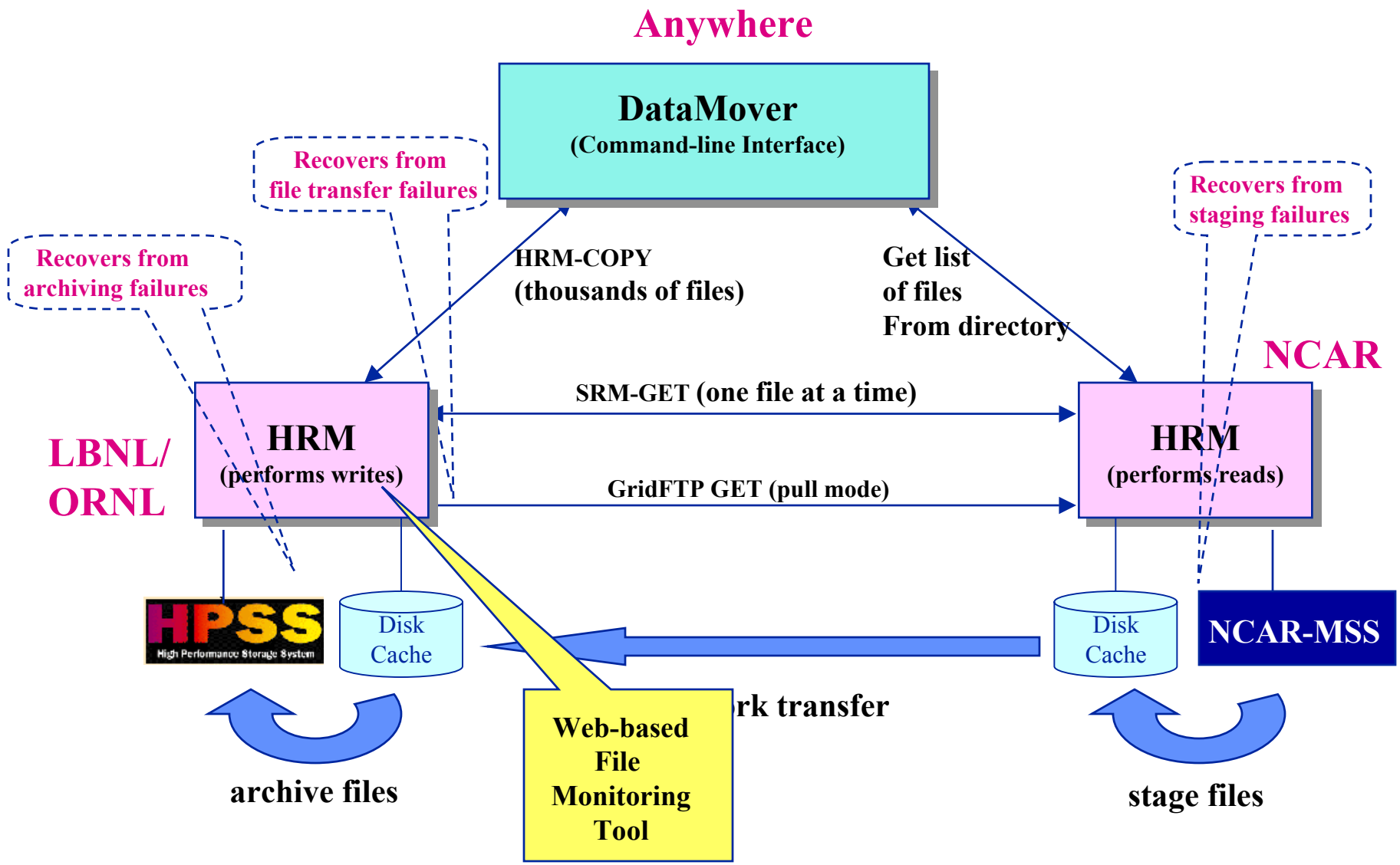
gridFTP
server

HRM
Storage Resource
Management

disk

HPSS
High Performance
Storage System

# DataMover: HRMs use in ESG and PPDG for Robust Muti-file replication



# Web-Based File Monitoring Tool

LBNL

Fermilab

**Shows:**

-Files already transferred

- Files during transfer

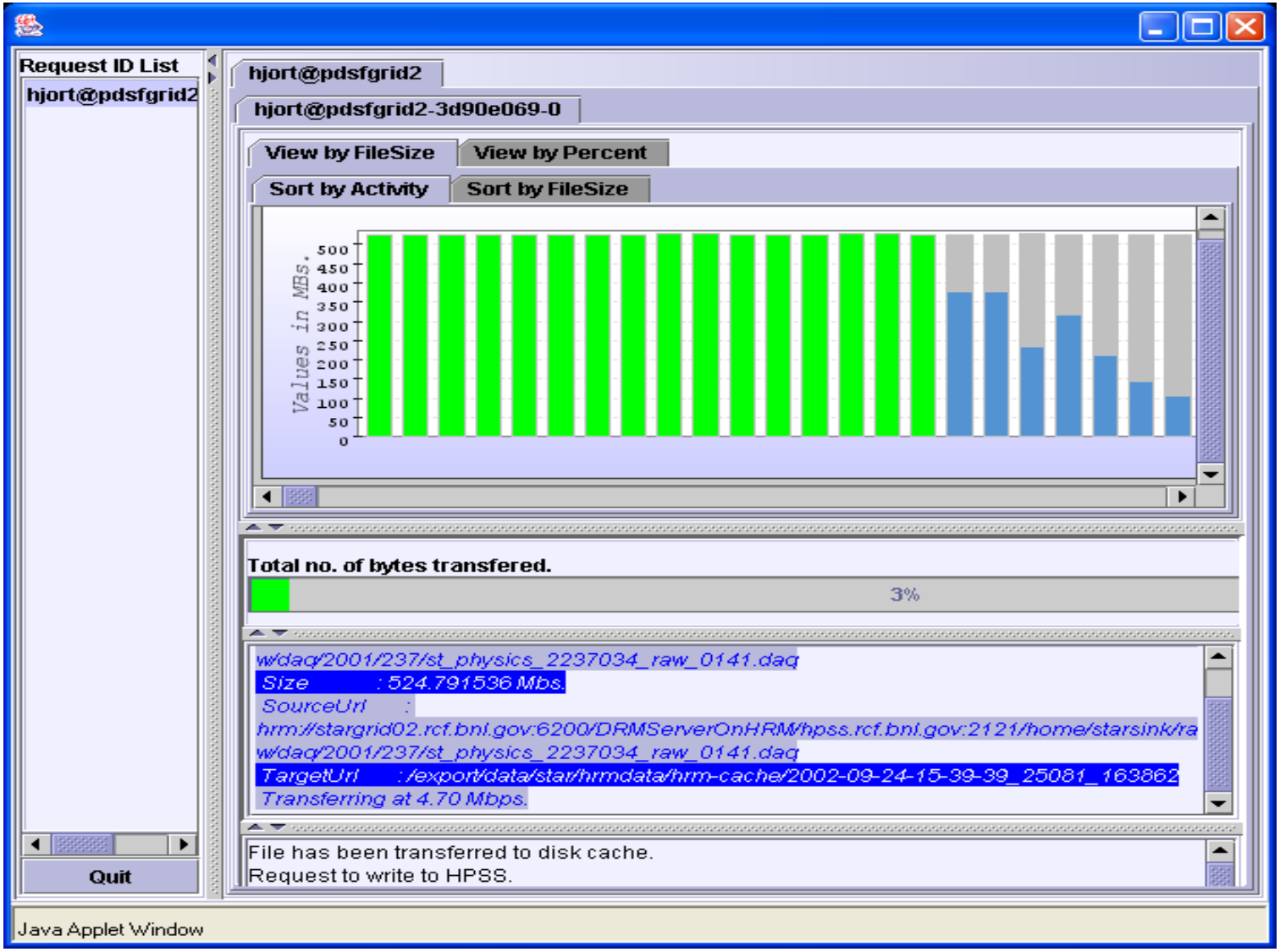
- Files to be transferred

**Also shows for each file:**

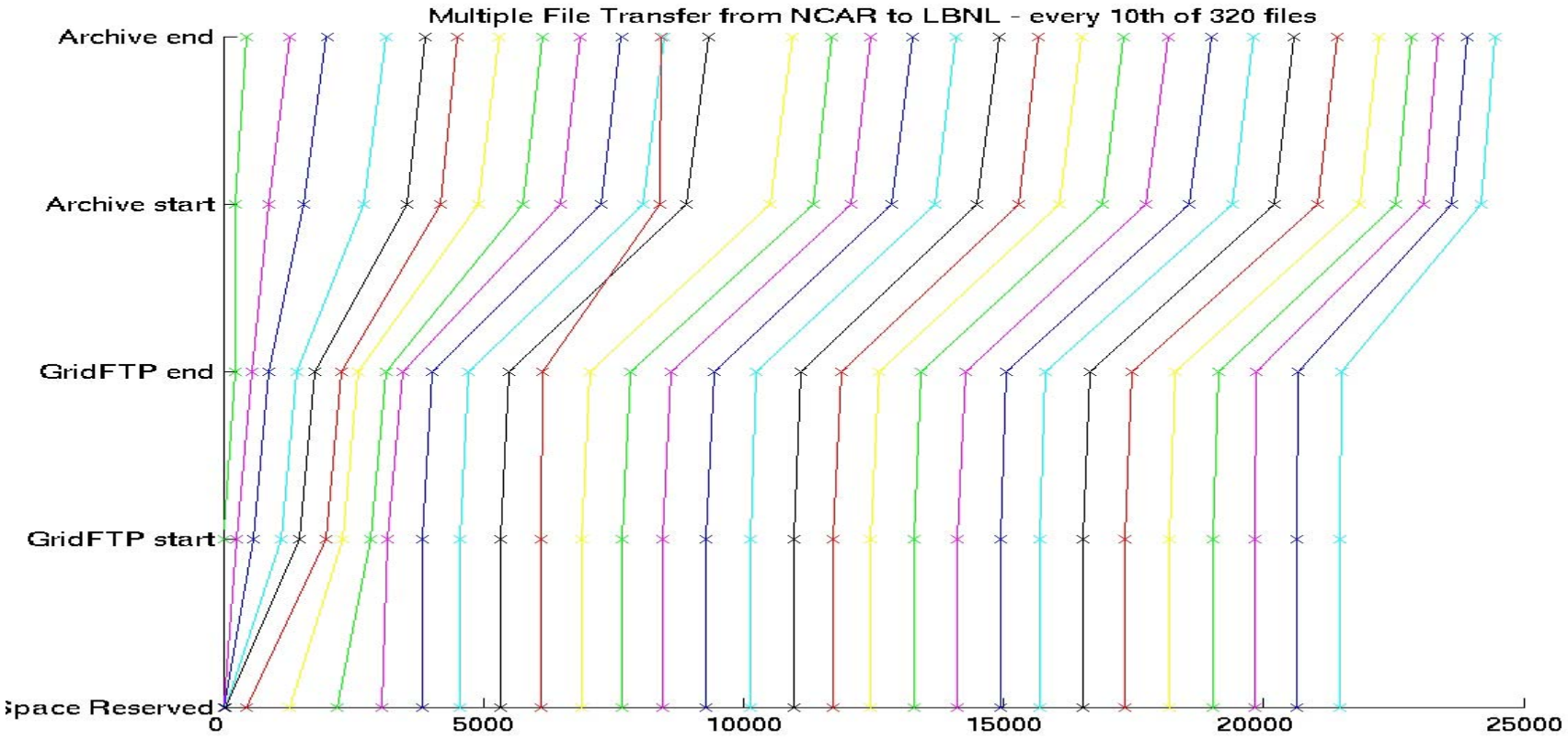
-Source URL

-Target URL

-Transfer rate

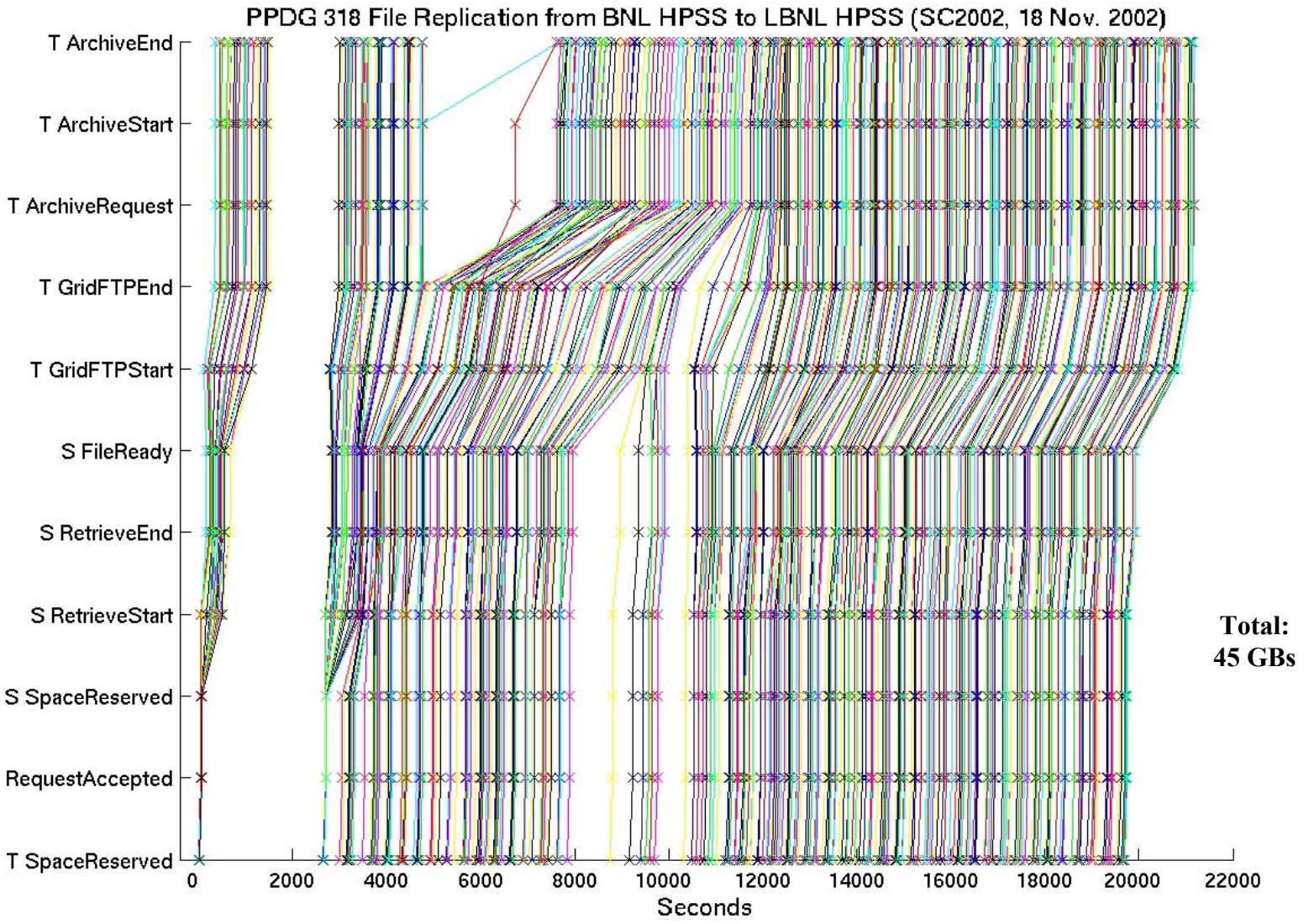


# File tracking helps to identify bottlenecks




**Shows that archiving is the bottleneck**

# File tracking shows recovery from transient failures




**Total: 45 GBs**

LBNL

# Statistics: end-to-end (staging, transfer, archiving)

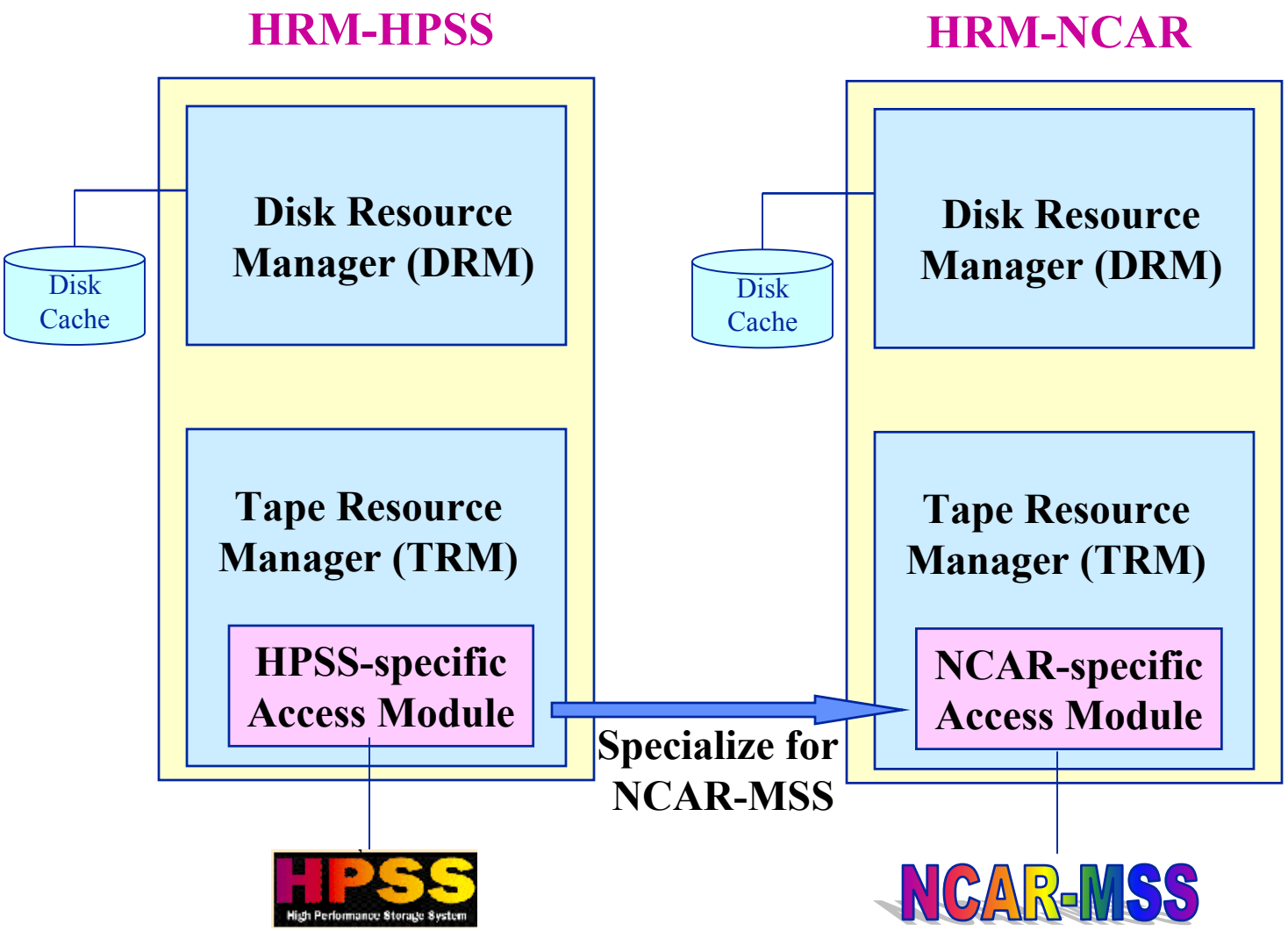
Fermilab

| Date | Number of Flies | Ave file size (MB) | Total Size (GB) | Total time (hours) | Ave transfer rate (MB/s) | Comments |
|---|---|---|---|---|---|---|
| 03/05/03 | 110 | 318 | 35.5 | 1.79 | 5.5 | NCAR-LBNL |
| 03/05/03 | 43 | 35 | 1.5 | .23 | 1.8 | LBNL-NCAR |
| 04/26/03 | 1010 | 50.7 | 51.2 | 6.94 | 2.04 | LBNL-NCAR |
| 04/25/03 | 1005 | 36.6 | 32.8 | 7.29 | 1.25 | + 14 hrs MSS failures |
| 4/28/03 | 505 | 34.5 | 17.4 | 2.9 | 1.66 | LBNL-NCAR |
| Desired ? | | | 1000 | 24 | 11.57 | |

# "Gridifying NCAR's MSS: Adapting HRM-HPSS for NCAR-MSS

LBNL

Fermilab



# standardization in US and EU



- **Ongoing work**
  - **Developing Standard SRM interfaces**
    - Particle Physics Data Grid (PPDG) project
      - LBNL, TJNAF, FNAL
    - European Data Grid (EDG) project
      - WP2 - data management
      - WP5 – mass storage (CASTOR)
    - Deployment
      - LBNL, BNL, ORNL, TJNAF, FNAL, CERN, (SE-England)
  - **Use of SRM by other agents**
    - Storage Resource Broker (SDSC) calling HRM to Stage files from HPSS
    - GridFTP invoking HRM

# New Features of SRM V2.1



- Space reservation services
  - Spaces and files: volatile, durable, permanent
  - Lifetime, action at end of lifetime
    - Volatile – SRM owned, files can be removed if space needed
    - Durable – files cannot be removed, but administrator notified
    - Permanent – can be removed by owner only
  - Support for “best-effort” reservation
- Directory services
  - Usual unix semantics
    - any type of files in directory
    - Extend with metadata of Volatile, Durable, Permanent
- Access control services
  - Support owner/group/world permission
    - Can only be assigned by owner
  - File sharing for read-only files
    - check with source for shared file permission
  - File sharing for updatable files
    - check with “master copy” for time of last update

# Future Plans - LBNL

- **Short term**
  - Extend Usage of DataMover to 5,000-10,000 files
  - Package DRM and DataMover to include with VDT
  - Develop a Replica Registration Service (RRS) for STAR experiment
  - Coordinate standards on Replica Management with SRMs
  - Deploy with Astrophysics – TSI project
- **Medium Term**
  - Develop a new DRM and HRM based on SRM v2.1 spec
    - Support space reservation
    - Support directory operations
    - Support native WSDL interface
  - Develop a version of DRM on top of NeST product from U of Wisc
    - Provide flexible support of durable and volatile space
    - Support best-effort space allocation

# Future Plans - LBNL



- **Longer Term**
  - Managing the cache: admission and replacement policies
    - Incorporate into DRM a "Grid-based" caching policies (results of basic research program)
  - Interface SRMs with Request Planners, Request Executers
    - Have SRMs provide dynamic storage availability
    - Experiment with space reservation negotiations
  - Provide Monitoring Packages with storage usage statistics
    - Via GLUE schema definition
  - SRM support of access authorization
    - based on Community Access Service (CAS)
    - Extensions to Akenti
  - Space accounting, and charging

# Overall View from HEP at Fermilab



- Support of diverse and data intensive experiments has entailed active use of tape based data sets
- HEP has a community which is knowledgeable, informed and interested in storage and data movement technologies
- HEP labs have massive, effective, and diverse local storage systems
  - Typically petabye archives. read-dominated tape plants
  - Large caches, commodity techniques
    - Scavenged disks on "worker nodes"
    - Commodity RAID servers
  - Have appropriate scalability, and work well
- HEP labs seek to standardize by converging to a "storage resource management" abstraction
  - Two components – transfer and management
  - Allow for system evolution and growth

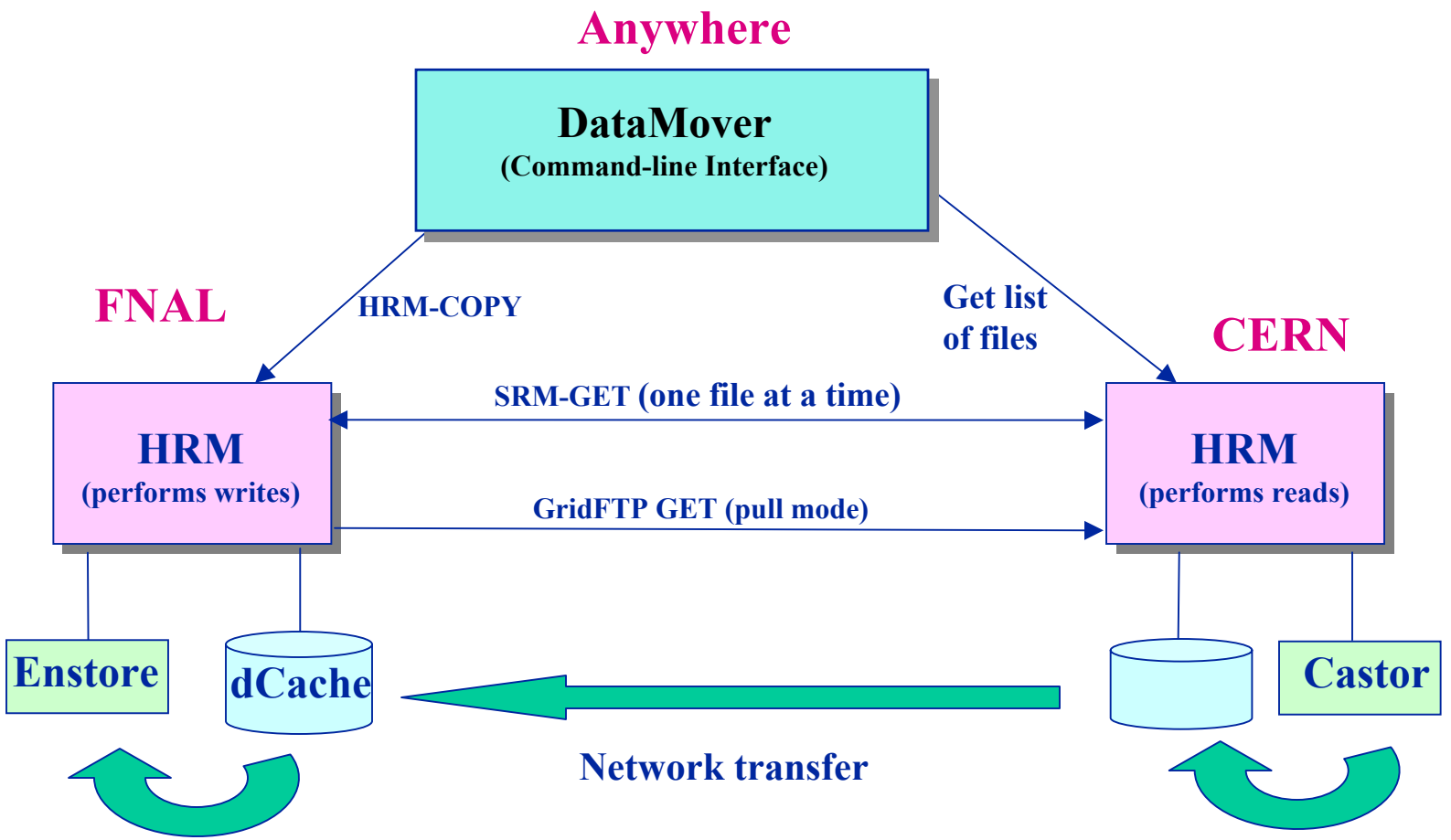
# LCG investigations at Fermilab



- Large HEP computing facilities have scalable, user-code file systems in production.
  - Partial file access, basically POSIX semantics
  - FNAL/CDF as an exemplar –
    - 100 TB/ide disk, 20 TB/day to analysis, 5TB/day faulted in from tape
- LAN Protocols are domain-specific, but very important:
  - CERN/ IN2P3 – rfio - accepted EDG protocol
  - FNAL/DESY/US CMS tier 2 – Dcache (Dcap protocol)
  - OTHER HEP grid sites use NFS
- FNAL/US CMS investigated using SRM to parameterize the "LAN side" protocols for LCG. (Large Hadron Collider Computing Grid)
  - Under consideration for LCG-2
  - Driven FNAL to provide C web services client for SRM V2

# US CMS Program of Work For Grid3





# SRM Status – CMS & LCG

- **CMS**
  - US CMS Goal: Grid3 demonstrations including technology verification/commissioning
  - for the US CMS DC04 milestone:
  - e.g.: Work on Robust Multi-file Replication
    - Data Streaming at 5% LHC scale out of CERN to Tier-1
    - Data Analysis between regional centers
- **LCG**
  - Statement of interest in SRM as a technology.
  - Ian Bird, Ruth Pordes

# SRM and Lattice Gauge



- FNAL, TJNAF/MIT, BNL are sites for dedicated SciDAC-funded Lattice Gauge computational facilities.
- However, The primary repository for Lattice datasets are:
  - NERSC
  - NCSA
- FNAL and JLAB are active participants in SRM discussion, and keep Lattice in mind.
  - Goal: support mirroring and local caching of these data sets.
- SRM parameterized I/O.
  - Accommodates NERSC->FNAL ingest via http.
  - Accommodates NSCA->FNAL ingest with extended GridFTP

# US/CMS



- The CMS user facility is participating in the 2003 CMS data challenge.
- We will begin using SRM for production data movement in the CMS data challenge this fall.
  - FNAL <-> CERN
  - FNAL <-> US prototype Tier 2 centers.
  - Will gain very valuable experience with sustained production.
  - First experience with sustained, daily interoperation with CERN.

# FNAL interests: SRM future (engineering)



- Have the standard WSDL, with convergence to grid practices.
- Have a body of client code for system integration, implemented in Java and C.
  - Promulgate SRM in the relevant middleware communities
- Have a library of Java classes useful for instantiating SRM services on other storage system
  - Allow the work in HEP related to storage elements to easily have standard GRID interfaces
- Promote SRM as a standard in LHC and LQCD systems.
- Continue collaboration/ support of U. Wisc/NeST.
- GGF/SRM participation.

# SRM Investigation Interests

- Study SRM as a way of sending other out of band commands and inquiries to a storage system.
  - Example: Inform replica catalog of file loss.
  - A derived interest is parameterized access to replica systems
- Understand SRM in the context of federated or Hierarchical grid base storage systems.
- Relate SRM to other storage models
  - BTeV distributed Permanent store.

# Summary: advantages of using SRMs



- Synchronization between storage resources
  - Pinning file, releasing files
  - Allocating space dynamically on “as-needed” basis
- Insulate clients from storage and network system failures
  - Transient MSS failure
  - Network failures
  - Interruption of large file transfers
- Facilitate file sharing
  - Eliminate unnecessary file transfers
- Support “streaming model”
  - Use space allocation policies by SRMs: no reservations needed
  - Use explicit release by client for reuse of space
- Control number of concurrent file transfers
  - From/to MSS – avoid flooding MSS and thrashing
  - From/to network – avoid flooding and packet loss
- Automatic “garbage collection”

# Summary – accomplishments



- Development of a standard
  - Mutli-file requests, file sharing, space management
  - New concepts: pinning, pin-lifetime
  - New concepts: durable files and spaces
- Coordination with GGF
  - We started the process - BOF
- Development
  - Five institutions
    - US (LBNL, Fermi, Jlab)
    - EU (CERN, Rutherford)
- Interoperation of Mass Storage & disk systems
  - HPSS, Enstore, JASMine, Castor, NCAR-MSS, dCache, unix-DRM
- Deployment
  - PPDG – BNL, LBNL, CMS, Lattice Gauge
  - ESG – ORNL, NCAR, LBNL, LLNL
- Publications

# Publications

LBNL

Fermilab

- **Book Chapter**
  - Storage Resource Managers: Essential Components for the Grid
    - *Arie Shoshani, Alexander Sim, and Junmin Gu*, in Grid Resource Management: State of the Art and Future Trends, Edited by Jarek Nabrzyski, Jennifer M. Schopf, Jan weglarz, Kluwer Academic Publishers, 2003.
- **Part of a book Chapter**
  - Storage Resource Management
    - The Grid: Blueprint for a New Computing Infrastructure, Edited by Ian Foster & Carl Kesselman
- **Conferences**
  - Storage Resource Managers: Middleware Components for Grid Storage,
    - *Arie Shoshani, Alex Sim, Junmin Gu*, Nineteenth IEEE Symposium on Mass Storage Systems, 2002 (MSS '02)
  - 2 presentations at CHEP 2003
- **News item for PPDG**
  - STAR/HRM achieves robust, effective, Terabyte-scale multi-file replication, http://www.ppdg.net/docs/news/news-update-star-hrm-25sep02.pdf
- **Tech reports**
  - 5 tech reports on the SRM design and SRM methods spec
    - sdm.lbl.gov/srm